Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 16:28:37
REFERENCE NO: 287

Author Names & Affiliations

- Michael Norman - San Diego Supercomputer Center; University of California, San Diego
- Frank Wuerthwein - San Diego Supercomputer Center; University of California, San Diego
- Shawn Strande - San Diego Supercomputer Center; University of California, San Diego

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

cross-disciplinary; data intensive cyberinfrastructure

Title of Submission

Toward an integrated, national cyberinfrastructure for data-intensive scientific discovery

Abstract (maximum ~200 words).

Investments by NSF have resulted in a highly capable, and diverse national cyberinfrastructure (CI) that allows researchers to answer the most pressing questions of our time. Despite these investments, a significant gap exists between the needs of the community and what NSF can fund, and CI is not as well integrated, nor as easy to use as it could be.
Recent successes in deploying CI provide a foundation for a new level of integration of research assets nationwide. We envision a federation of national CIs that seamlessly connect research at campuses nationwide with national supercomputing centers and commercial clouds, enabling collaboration among scientists across disciplines and institutions on resources they own, share, have allocations or buy time on.
At the root of success is the conviction that through partnerships between domain scientists and centers like SDSC we can today deliver an integrated, national cyberinfrastructure that serves NSF's research and education mission, and impacts science in the nation across institutional, disciplinary, and agency boundaries. We encourage NSF to seize this opportunity to maximize the impact of its prior investments by providing the leadership necessary to achieve a system of collaborating national CIs that allows integration of all research assets nationwide.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Since 1981, when the NSF's supercomputing center's program was conceived as part of a larger effort to build a national high performance network for advanced research and education, NSF has continued to make significant investments in a broad spectrum of computing,

software, networking, and expertise. Programs like the Extreme Science and Engineering Discovery Environment (XSEDE), Software Infrastructure for Sustained Innovation (SI2), and Campus Cyberinfrastructure – Data, Networking, and Innovation (CC*DNI) gives thousands of researchers access to the most advanced CI and expertise available and enables breakthrough discoveries across virtually every domain of science. These, and other NSF programs, have allowed centers like SDSC to innovate and create comprehensive CI solutions that significantly broaden the impact of NSF investments beyond what is possible through any single program or by any one institution. Recent examples include:

? The rapid growth and adoption of science gateways[i] gives domain specific communities a way to interact with CI through simple web interfaces, and allows them to focus on productive science;
? Virtualization technology makes it possible to deploy community-specific software on traditional HPC systems. For example, the Open Science Grid was recently integrated with SDSC's Comet[ii] system, and allows projects such at LIGO and LHC to carry out high-throughput computing as part of their data analysis. This work also paves the way for integration with the commercial cloud;
? The Pacific Research Platform[iii], a partnership of the 10 UC campuses, Caltech, USC, Stanford, the University of Washington, and others to create a regional high-performance network (1000 times faster than current inter-campus networks), providing a foundation for a broad range of data-intensive applications in areas such as accelerator particle physics, astronomical telescope survey data, gravitational wave detector data analysis, and cancer genomics.
We believe these successes can be traced to four major factors:
1. Technology maturity. CI, and particularly high speed networking, has reached a level of performance and stability such that it is now possible to build, integrate, and operate distributed systems and services in ways not previously possible;
2. Science as the driver. A well-developed community of researchers with a specific need came first, with CI deployed on their behalf;
3. Campus bridging. Significant investment at the campus level [iv] has resulted in research computing infrastructure that provides the foundation to bridge from the campus to national resources [v];
4. Sustained leadership. Sustained leadership from centers like SDSC has made it possible to build, maintain, and grow the human expertise and networks of collaborators needed to design, deploy, and operate complex CI on behalf of the science community.

These four factors strongly indicate that we are at a propitious moment in the evolution of CI such that if a program were announced that funded science-driven CI integration activities organized around today's scientific challenges and national priorities the community is prepared to respond coherently and with purpose.

[i] In late 2015, the number of users accessing NSF's XSEDE resources via science gateways exceeded the number doing so from traditional command line.
[ii] http://www.nsf.gov/news/news_summ.jsp?cntn_id=136638
[iii] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1541349&HistoricalAwards=false
[iv] NSF's CC-NIE, CC-DNI, CC *, and the Major Research Instrument (MRI) program have been important sources of funding for CI at the campus level. http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5260
[v] NSF Advisory Committee for Cyberinfrastructure Task Force on Campus Bridging. Final Report, March 2011.
Available from: http://www.nsf.gov/od/oci/taskforces/TaskForceReport_CampusBridging.pdf

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

Discipline-led CI in partnership with national centers

We encourage NSF to bring forth a major solicitation that invites leading researchers in a data-intensive discipline to join with centers like SDSC, to deliver a disciplinary, national-scale integrated cyberinfrastructure on behalf of a major research agenda. Elements of such a program would include:
? Being led by top researchers in a data discipline who are working together with supercomputing, networking, visualization and data storage;
? Data sources and repositories which are distributed on a regional and/or national scale;
? Building on top of programs like CC*NIE at each of the campuses, interconnected through regional networks to form a common

data/compute arena, all interconnected at 10-100Gbps and beyond;
? In addition to solving a single disciplinary science goal, require a strategy for engagement with all of open science to guarantee broader impact of the CI developed beyond the immediate boundaries of the primary stakeholders;
? Inviting NSF directorates to a joint, cross-agency call; and,
? Funding 2-3 projects at $10-20M/each over 5 years, with the potential for a 5-year follow-on.

Integrated CI will accelerate scientific discovery and leverage NSF investment

The examples noted above are strong indicators that the proposed program would have a profound impact on the role of CI in scientific discovery. We expect it would:
1. Provide direct and measurable impact on targeted data-intensive research projects;
2. Enhance partnerships across the NSF directorates, and demonstrate the value of CI to NSF's mission;
3. Allow for the evolution of SI2 and DIBBs to encourage interoperability within both programs and better leverage NSF investments to support a diverse but interoperable national CI for data exploration;
4. Provide a blueprint for how NSF investments across the CI ecosystem can be combined into an integrated CI on behalf of the scientific community.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

There are several areas where progress is needed if integrated CI is to achieve its full potential of meeting the needs of the research community. These include:

1. There is a serious shortage of expertise needed for developing, supporting, and innovating CI on behalf of the research community. This shortage is exacerbated by competition with the private sector for professionals across software, systems, storage, and networking fields. It is the integration of these fields that is of most value to building CI, and thus it is imperative that NSF, in partnership with the community, develop programs that target these interdisciplinary skills and attract those to the workforce who have the interest and potential to build the next generation of integrated CI.

2. The emergence of commercial cloud computing is both an opportunity and a challenge. While the availability of low-cost computing cycles is a good fit for some communities and should, in the long run, lead to cost efficiencies for many, the work that is needed to embed this into a usable CI fabric remains misunderstood. Investment is needed in integrating software, workflows, and systems so that the research community experiences the cloud as part of an integrated ecosystem and can focus on the conduct of productive science and less on where they compute.

3. Given the significant expertise, facilities, and other infrastructure costs of operating large scale CI, and more importantly, the hidden cost of researchers having to pick up and move their data, tools, and research when a new system comes on line at a new center, there must be models of sustainability that allow this infrastructure to persist for at least 10 years, and support for researchers to transition from one system to the next.

4. Emerging technologies and software, such as many core, quantum and neomorphic computing, machine learning, and a plethora of new memory and storage technologies will place significant new requirements on scientific application developers who need these tools to deal with ever increasing computational and data analysis challenges. While some of this is being addressed through, for example, DOE programs, the vast majority of researchers (aka the "long tail of science") will require training to understand how and when to make use of these technologies. Likewise, CI providers need to investigate these technologies and understand how to integrate them into usable CI. Thus, NSF should support exploratory programs that are intended to bridge from research in these technologies to how they can be used by end users and CI providers.

Consent Statement

- "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to

display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode)."